# Gathering high quality information on landslides from Twitter by relevance ranking of users and tweets

Aibek Musaev, Qixuan Hou
Georgia Institute of Technology, Atlanta, Georgia
{aibek.musaev, qhou6}@gatech.edu

***Abstract*—Social networking platforms are increasingly used to report or pass along news and other valuable information. Their use rises especially during emergency situations and can be monitored for the analysis of adverse events, such as disasters. In this paper, we provide an overview of a comprehensive disaster information system using social networks with landslides serving as an illustrative example. We briefly describe each of the steps involved and focus on the classification and ranking steps that determine the relevance of individual messages and groups of messages to landslides. We introduce the concept of "relevant" and "irrelevant" virtual communities of users and compute their influence in each of them. This allows us to improve the existing relevance ranking formula by taking into account not only the semantics of the messages posted by users, but also the users' influence and the amount of their activity in these communities to improve the quality of the collected information on landslides. The resulting system achieves 0.936 F1-score in classifying individual messages and 0.941 F1-score in relevance ranking of the events.**

***Index Terms*—Social networks, text classification, landslide, PageRank.**


Fig. 1. Early image of the mudslide in Oso, Washington

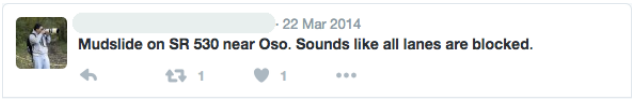


Fig. 2. One of the earliest reports of the mudslide in Oso, Washington

## I. Introduction

Social networking platforms have experienced remarkable growth in recent years. For example Twitter, which is a popular microblogging social network, has more than 240 million active monthly users. A typical daily activity on Twitter is more than 500 million tweets, which corresponds to an average of 5,700 tweets per second [1]. The use of social networks rises even higher during disasters [2]. Not only emergency response agencies and news sources, but also regular users spread information in safety-critical situations [3].

See Figure 1 for an example of one of the earliest tweets containing an image of a recent disaster. It shows the fatal mudslide in Oso, Washington that occurred on March 22, 2014. A portion of an unstable hill collapsed covering an area of approximately 1 square mile and killing 43 people[1]. This disaster caused a spike of activity in social networks in the aftermath of the event. One of the earliest reports about it came from a local Twitter user as shown in Figure 2.

The work on detection and analysis of disaster events enables researchers to take the next step, which is the prediction of disasters. If we learn how to predict disasters then we can potentially save many lives and reduce the impact of future disasters on society. Recently, floods in north and central China have killed at least 150 people, with many still missing and hundreds of thousands forced from their homes[2]. The many casualties from this horrific disaster prompted people to protest against the government for failing to warn them of the flooding. If people had been warned of the flooding then they would have been able to escape the affected area beforehand.

But before we learn how to successfully predict disasters based on the data from multiple sources, such as social networks, we need to understand how to collect high quality information about disasters first. In this paper we describe our work on content filtering and relevance ranking of Twitter users and tweets on landslide events involving the movement of soil.

[1]http://snohomishcountywa.gov/ArchiveCenter/ViewFile/Item/3913

[2]http://www.bbc.com/news/world-asia-china-36873902

## II. Overview and Related Work

Given a set of messages collected from social networks using disaster related search keywords, we are interested in detecting high quality messages related to disasters. Specifically, using the data collected from Twitter based on search keywords *landslide*, *mudslide*, and *rockslide*, we want to detect landslide events involving the movement of soil. This problem can be broken down into several subproblems as follows:

- Collecting messages containing search keywords
- Geotagging of individual messages
- Classifying of individual messages by relevance to disasters
- Grouping of geotagged and classified messages to events
- Ranking events by relevance to disasters

We briefly describe each of these steps next, where the foci of this work are the classification and ranking steps that measure the relevance of individual messages as well as groups of messages (or events) to disasters.

**Collecting messages containing search keywords**. At the moment, multiple popular social networking platforms allow programmatic access to their datasets based on search keywords, including Twitter and YouTube. Twitter has the most advanced set of APIs among the social networks we considered by providing two types of access that are based on either push or pull architectures. We implemented a push-based Streaming API client, which provides low latency access to Twitter's global stream of tweets[3].

In this project, we use a public dataset available here[4]. It contains data from multiple social networks, such as Twitter, which was collected during 2014 based on multiple search keywords related to landslide events, namely *landslide*, *mudslide*, and *rockslide*. In addition to the original data provided by Twitter, the messages had been automatically geotagged and manually annotated with respect to their relevance to landslide as a natural disaster as described below.

**Geotagging of individual messages**. We treat disasters as events that are defined by their spatiotemporal features, which is why we need to retrieve not only the timestamp of each message collected from social networking platforms, but also the geographic location associated with it. However, less than 0.42% of all tweets in their dataset are geotagged [5]. This is a very important research challenge, which is outside the scope of this paper. In this work we use a dataset of tweets containing landslide keywords that had been geotagged already. The geotagging process is described in Section VI-A.

**Classifying of individual messages by relevance to disasters**. Due to ambiguity of the used search keywords, the data collected from social networks may be either relevant or irrelevant to our topic of interest, which is disasters. Sakaki, et al. treat users as sensors and observe their tweets to detect earthquakes [7]. However, social users are an example of very noisy sensors compared to physical sensors. For example, *shaking* may be referring to someone shaking hands instead of an earthquake. And trivial approaches based on identification of sudden bursts of social activity containing the search keywords may lead to a large number of false detections [8]. Sakaki, et al. propose to use machine learning classification to automatically label tweets as either relevant or irrelevant to disasters based on the classification model that is learned using an annotated training dataset [7]. A critical part of this process is the decision of how to convert textual information such as tweets into numerical format expected by a classifier algorithm, where each number represents a feature of the original tweet. This part of the classification process is also known as feature generation. We describe our approach to classification and feature generation in Section III.

**Grouping of geotagged and classified messages to events**. Once the messages are geotagged and classified, they are then grouped into events based on their spatiotemporal features. Musaev, et al. propose to represent the surface of the Earth as a grid of cells [9]. Each geotagged message is mapped to a cell in this grid based on the item's geographic coordinates retrieved during the geotagging process. This cell-based approach is studied further to improve the accuracy of the geotagging process through a composition of clustering methods [6], [10]. We apply the proposed approaches by grouping the tweets within each month by their cells, which are computed based on their geographic coordinates as described in Section VI-A. Our assumption is that there can be only one disaster within a month per cell. In other words, we define an event as a group of messages that are mapped to the same cell within a 30-day window. See more information about it in Section VII.

**Ranking events by relevance to disasters**. After the previous steps are executed, the system obtains a list of potential disaster events represented as non-empty cells with a set of classified tweets that are mapped to each of those cells. Our objective is to rank the non-empty cells based on their relevance to disasters, such as landslides. In other words, instead of producing a boolean decision of whether a landslide occurred in each cell or not, we want to assign a probability of landslide occurrence to each cell. This value can be computed as the relevance to disasters and can be used to rank those cells by the likelihood of disaster occurrence.

Musaev, et al. propose a Bayesian model as the relevance ranking strategy as follows:

$$P(w|x) = \sum C_i \frac{N_i^x}{N_i^x + 1}, \tag{1}$$

where $C_i$ denotes the normalized prior F-measure of source $i$ from historic data, $N_i^x$ denotes the number of items from source $i$ in cell $x$ indicating that a landslide $i$ occurred in the area covered by cell $x$ [9].

This approach is improved further in Equation (1) by taking into account not only the messages that are classified as relevant to a disaster, but also the irrelevant ones:

$$P\left(\omega|x\right) = \sum_i R_i \frac{\sum\limits_j POS_{ij}^x - \sum\limits_j NEG_{ij}^x - \sum\limits_j STOP_{ij}^x}{\sum\limits_i N_i^x}, \tag{2}$$

where $POS$ denotes relevant or positively classified tweets, $NEG$ denotes irrelevant or negatively classified tweets, and $STOP$ denotes the tweets containing stop words or stop

[3]https://dev.twitter.com/streaming/overview

[4]https://grait-dm.gatech.edu/resources/

phrases [11]. This formula implements the idea of penalized classification, such that for each cell the majority label is accepted and only cells whose majority label is "relevant" are considered.

However, the relevance ranking formula shown in Equation 2 has several critical shortcomings. Notably, it ignores the number of items mapped to a cell. Consider the following scenario as an example. Let us assume that there is a cell A with 1,000 positively classified tweets and a cell B with only 1 positively classified tweet. We should have a much higher confidence that there was a disaster in cell A compared to cell B. However, based on the formula in Equation (2) both cells would be assigned the same score, which is equal to $R_i$, i.e. the normalized prior F-measure of sensor $i$, such as Twitter.

Furthermore, this formula ignores a user's influence in the subject and assigns the same weight to each tweet, which is equal to 1. In other words, regardless of a user who posts a tweet, all tweets receive the same vote of confidence from the system. However, social users have varying amounts of influence in different subjects [12].

A post by an authoritative source, such as the US Geological Survey agency (@USGS), should have a higher score than a post by an ordinary user with no history of interest in this subject and no connections with such users. In addition to these two extreme cases, there are also users that may be interested in multiple subjects, including relevant and irrelevant topics to disasters. For example, news sources, such as CNN, are influential users, but they frequently post content which may be either relevant or irrelevant to disasters. In such case, the prominent news sources will have a high influence in the topics that are relevant to disasters, such as landslides, and the topics that are irrelevant to landslides, such as politics and sports.

The messages are classified by the system as either relevant or irrelevant to disasters. Based on the relevance of the messages that users post, we propose to build two virtual communities:

- "relevant" community: the users who post messages that are classified as relevant by the system or expressed interest in those messages;
- "irrelevant" community: the users who post messages that are classified as irrelevant by the system or expressed interest in those messages.

Social networking platforms allow users to express their interest in a given message, e.g. by sharing it (Facebook) or retweeting it (Twitter). The dataset that we use in this paper contains messages collected from Twitter as described in Section VI-A. Twitter's Retweet feature enables users to share a post that they are interested in with their followers. A retweet not only expresses a user's interest in the subject of the original tweet, but it can also serve as an instance of a directed relationship between users based on topical interest.

Note, that the fact that user A retweets a message by user B represents a citation of user B by user A. Therefore, to determine a user's influence in the "relevant" and "irrelevant" virtual communities, we propose to apply the PageRank algorithm [13], such that the users serve as the nodes and retweets as the directed edges or links between nodes. Also note, that a user may have two scores corresponding to her influence in the "relevant" and " irrelevant" communities as she may be interested in both subjects.

Finally, we propose an improved formula to rank events by relevance to disasters, which takes into account the number of messages associated with each event and the influence of the users that post or forward those messages – see Section V for more information.

## III. Tweets: Classifying Tweets by Relevance to Disasters

The data collected from social networks frequently contain noise due to the use of polysemous search keywords [7]. By noise we mean the messages whose topics are irrelevant to the meaning of the search keywords that we are interested in:

- "landslide" denoting an overwhelming victory in politics: I don't think LANDSLIDE is the right word to describe how this election is shaping up. I'm thinking GALACTIC SHIFT.
- "landslide" denoting an overwhelming victory in sports: Brady got robbed, still the best QB in the league by a landslide, cant wait till superbowl 51
- "Landslide" as the name of a popular song by the 70s rock band: my all time fave fleetwood mac song :) #NowPlaying Landslide by Fleetwood Mac i've been afraid of changes coz i b...

For the purpose of disaster detection, we want to label the messages coming from social networks as either relevant or irrelevant to the type of disaster we aim to detect, such as landslide events. Moreover, we want to automate this process, such that the messages retrieved from social networks are labeled automatically with the minimum amount of human involvement while maintaining high accuracy.

The basic approach for automatic labeling of irrelevant tweets is based on the presence of stop words and stop phrases in a tweet's text, such as *fleetwood* and *election* or the lyrics from the "Landslide" song by Stevie Nicks from Fleetwood Mac: "...and I saw my reflection in the snow covered hills...". However, after applying this trivial approach there are many tweets that remain unlabeled.

For the remaining tweets we apply a machine learning technique called classification that is specifically designed for this automated task. An expert manually labels only a part of the data for training purposes, which is a one-time operation. Given such dataset with preassigned labels, called ground truth, a classifier algorithm builds a model of how to interpret the data, such that it can automatically generate a label for, i.e. classify, an unseen data item. In our case we consider the following classes: either "relevant" or "irrelevant" to landslide as a disaster. This is an example of binary classification as there are two classes involved.

For evaluation purposes, another dataset is needed with correct labels in order to check the accuracy of the proposed classification model. Although the cross-validation method can be applied for this purpose [14], in this paper we use real world data collected from Social Media during the full year of 2014 – see Section VI-A for the description of the evaluation dataset.

Note, that classifier algorithms expect numerical data, whereas our dataset contains messages, i.e. textual data, collected from social networks. Thus, we need to convert each message to a list of numbers also known as a vector. We want this conversion to be a good representation of the text, such that the classification model generated using these vectors, had a high accuracy when predicting labels for unseen messages. The numbers in the vectors serve as features of the messages for classification purposes and are consequently called classification features. Based on our experience, the choice of features used for conversion is crucial for building a robust classification model that can compute labels with high accuracy.

Over the course of this project we experimented with various feature generation methods and observed varying results. The standard baseline approaches, such as Bag-Of-Words (BOW) [15] and statistical features [7], produce the worst classification results. Explicit Semantic Analysis (ESA) based methods [16] result in strong performance, however they require relatively large vectors with, for example, 2,400 dimensions that take longer to compute, especially given large datasets.

The authors of the recently proposed Continuous Bag-of-Words and Skip-gram model [17] published pre-trained word vectors with 300 dimensions. We build a centroid-based classifier [18] using these vectors as features as demonstrated in Section VI-B. Based on our evaluation, an SVM classification model based on these features has a high accuracy while being reasonably fast to compute.

For actual classification we use Weka, which is a suite of machine learning software developed in Java [19]. Weka implements multiple classification algorithms, including Naïve Bayes, C4.5 (a decision tree based algorithm), Random Forests, Logistic Regression, and Support Vector Machines (SVM). Due to paper limitations, we do not include the results of comparison between these algorithms. We select SVM as the classification algorithm due to its robust performance while being reasonably fast to compute based on the experiments.

As can be seen in Section VI-B, the proposed classification model demonstrates good performance, but there are still tweets that are misclassified. This can lead to either missed or falsely detected events. That is why the relevance ranking strategy should take into account not only the computed label of each message, which may be faulty, but also the past history and the influence of the users who post those messages. We describe our approach that takes into account both the computed labels of the messages and the influence of the users who post them in Section V.

## IV. Users: Ranking Users by Relevance to Disasters

In the previous section we described the analysis of the relevance of individual tweets to our topic of interest, which is landslides. Specifically, we apply a machine learning technique called classification. However, even though the classification performance is robust as can be seen in Section VI-B, there are still tweets that are misclassified. But the tweets do not exist in a vacuum. They are posted by real users, who have certain interests that are shared by the members of the corresponding virtual communities. Moreover, the users have varying degree of influence and activity in a given topic. As an example, a tweet posted by an authoritative source, such as USGS, should have a higher value than a post tweeted by a recent user with no history of interest in this subject or connections with users interested in this subject.

When we analyze the messages posted by Twitter users in our dataset, we can observe that there are three types of users with respect to their interest in landslides:

- *relevant*: the users who post messages that are mostly relevant to landslide as a natural disaster;
- *irrelevant*: the users who post messages that are mostly irrelevant to landslide as a natural disaster;
- *combination*: the users who post messages that may be relevant and irrelevant to landslide as a natural disaster.

Here are a few examples of social users who typically post messages related to disasters only:

- @*USGS*: Landslide Preparedness... Landslide Warning Signs. What to do before, during and after. http://on.doi.gov/1Ogc6ek
- @⟨*private user*⟩: I added a video to a @YouTube playlist http://youtu.be/H0zOfyMYRLM?a South China Landslide: No casualties reported in Hunan Province
- @*ERrisk*: #NaturalDisastersNews - Floods ravage Assam; two dead in landslide in Guwahati http://ift.tt/29H5p8Z

The following are examples of social users whose messages are virtually all about topics that are irrelevant to landslides, such as politics, sports or entertainment:

- @⟨*private user*⟩: Starting to think a Tory maj is disastrous for the SNP. Landslide or not, they'll have all the clout of the pre-2010 Lib Dems. ie: NOT MUCH.
- @*KPIsports*: @RockChalkBlog @CrimsonBlueKU KU has No. 1 SOS by a landslide. The gap between KU & No. 2 Florida is same as gap between SOS No. 2 & No. 97.
- @⟨*private user*⟩: here's my video of landslide from last night! <3 perfect audio i think! xo http://www.youtube.com/watch?v=8VUPWdUGVw4

  ...

The last category of users posts messages about topics that may be relevant and irrelevant to landslide as a disaster, including various news agencies:

- @*NYMag*: How much should we make of the New Hampshire polls showing a huge Kasich surge and a Sanders landslide?
- @*10News*: Landslide swallows house, forces dozens of residents to evacuate watch #WorldNewsTonight on ABC10 at 6:30
- @*rssworldnews*: Yahoo News : Japan's ruling bloc wins landslide in upper house election: exit polls

Recently, Weng, el al. propose to apply a TwitterRank algorithm, which is based on PageRank [13], to find topic-sensitive influential users [20] . Instead of using the "following" relationships in Twitter due to homophily, they identify the topics that users are interested in based on topic modeling.

Specifically, they apply the Latent Dirichlet Allocation (LDA) model, which is an unsupervised machine learning technique.

Our approach is based on the supervised machine learning classification technique instead. In our model, a given user may be active in the community consisting of relevant users, or in the community consisting of irrelevant users, or both. We propose to build such communities based on the relevance of tweets that users post and the relationships between them. The relevance is determined using machine learning classification and the relationship between users is based on their retweet behavior.

Our dataset consists of tweets containing the mentioned search keywords. Each tweet is classified by the system as either *relevant* or *irrelevant* to our topic of interest, which is landslides. Each tweet also contains information about the user who posted it and whether it is an original tweet or a forwarded tweet. A tweet can be forwarded to all of the user's followers, in which case it is known as a retweet. Retweets are often used to pass along news or other valuable information on Twitter.

We propose to build virtual communities of users based on the "retweet" relationships. Note, that this information is available in the tweets and does not require separate API calls, which would be subject to API limits[5]. In our case we have two communities consisting of users who post or forward tweets that are "relevant" or "irrelevant" to landslide as a disaster. And as we point out earlier, there are users who exist in both of these communities, such as news sources.

Given a community of users and a set of links between them based on the retweet relationships, we apply the PageRank algorithm to compute the influence of users in the corresponding communities. PageRank is a link analysis algorithm which counts the number of backlinks, i.e. the incoming links, to a node to determine a rough estimate o how important it is. We treat users as nodes and treat the count of a user's retweeted tweets as a vote of support. The more retweeted tweets the users have, the higher ranking score they will get. Therefore, the final ranking score is based on the user's influence.

We calculate the ranking scores in two communities, the "relevant" and "irrelevant" to landslide as a natural disaster. If a user is active in both communities, such as news sources, then she will have two scores computed for her.

Note, that we take into account not only a user's influence in general, but also her influence with respect to landslide as a disaster. For instance, the proposed strategy produces high relevance ranking scores for authoritative news sources, such as Wall Street Journal (*@WSJ*), BBC (*@BBCWorld*, *@BBCNewsbeat*), and Yahoo News (*@YahooNews*). At the same time, it generates high scores for local users, such as User1, User2, and User3. Although User1 is not a famous or influential user in general, she gets a high relevance ranking score since she actively works as a journalist and posts tweets about various events in her city. Note, that we replace the actual screen names with anonymous labels for these regular users to preserve their privacy.

[5]https://dev.twitter.com/rest/public/rate-limiting

## V. Events: Ranking Events by Relevance to Disasters

For each cell we propose to compute the relevance ranking score as follows:

$$\begin{aligned} P(\omega|x) = &\sum_i RelevantRank(POS_i^x) \\ &- \sum_j IrrelevantRank(NEG_j^x) \\ &- \sum_k IrrelevantRank(STOP_k^x), \end{aligned} \tag{3}$$

where for each cell $x$, $i$ is the number of positively classified tweets that are mapped to that cell, $j$ is the number of negatively classified tweets that are mapped to the same cell, and $k$ is the number of tweets containing a stop word or a stop phrase that are mapped to that same cell. Note, that the old formula in Equation (2) treats all tweets equally, such that each tweet gets the same weight of 1. The new formula in Equation (3) improves it by taking into account the relevance ranking of each user as the weight for that user's tweets.

Also observe that both the negatively classified tweets and the tweets containing stop words or phrases, get the user's rank from the "irrelevant" community as in both cases the tweets discuss an irrelevant to landslide as a disaster topic. Whereas if a tweet is classified as "relevant", then the user's rank from the corresponding community is used.

Since the formula uses the users' ranks instead of considering all users equally, influential users, such as *@BBCNews* or *@WSJ*, affect the score more than ordinary users. This also means that the cost of a classification error for such users is high, that is why it is important to use an accurate classification algorithm. Also, relevant users, such as *@climateprogress*, affect the cell's score more than ordinary users.

See the results of the experimental evaluation of the proposed approach in Section VI-D.

## VI. Evaluation Using Real Data

### A. Dataset Description

For evaluation purposes, we use the annotated dataset for landslides available here[6]. Specifically, we consider the evaluation dataset containing messages collected from Twitter during 2014 based on the keywords related to landslide disasters, such as *landslide*, *mudslide*, and *rockslide*. Conveniently, it is broken into separate files for each month.

The dataset includes only geo-tagged items. The items are geo-tagged based on the mentions of geographic places in them. Geographic terms are retrieved using the Stanford NER library [21] as location entities. Locations are then converted to geographic coordinates (i.e., latitude and longitude values) using Google Geocoding API [22]. Finally, the events are defined by their spatiotemporal features, such that an event's location is a cell, whose row and column values are computed as follows [10]:

$$row = (90^\circ + N)/(2.5'/60') = (90^\circ + N) * 24, \tag{4}$$

[6]https://grait-dm.gatech.edu/resources/

$$column = (180^\circ + E)/(2.5'/60') = (180^\circ + E) * 24, \quad (5)$$

where $N$ and $E$ are a tweet's latitude and longitude coordinates, respectively. The idea behind this cell-based approach is that the surface of the Earth is covered with a grid consisting of cells and the tweets are mapped to cells in this grid based on their geographic coordinates. The size of a cell in this grid is roughly equal to 2.3 miles by 2.3 miles.

Note, that since $N \in [-90^\circ, 90^\circ]$ and $E \in [-180^\circ, 180^\circ]$, then there are 4320 x 8640 $\approx$ 37 million such cells in total, which is a huge number to consider for any algorithm. However, we only need to analyze non-empty cells, which is in the order of several thousand cells per month for this dataset of landslide events.

In addition to the automated geo-tagging using the described cell-based approach, each tweet in the dataset is manually labeled as either *relevant* or *irrelevant* to landslide as a natural disaster. A human annotator analyzes the tweet's text to determine its relevance to a disaster. And if it contains a URL, then she looks at the URL to confirm the candidate item's relevance to landslides [23].

Thus, the overall objective of the system is to assign the correct label (*relevant* or *irrelevant*) to each non-empty cell during each month.

### B. Evaluation of Tweet Classification

In this experiment we compare the performance of the proposed Word2Vec based classification method versus a baseline approach. We select the standard Bag-Of-Words (BOW) algorithm as the baseline method.

*Baseline approach (BOW)*: BOW is a common baseline method, which treats each document as a bag of words. Using the training dataset we select the most frequently used words as the features of the BOW model, excluding stop words. Specifically, we set N=2,400, which is eight times larger than the number of dimensions in the proposed Word2Vec based method described below. Using these terms as features we select a binary representation of the messages in the training and evaluation datasets based on the presence of each feature as the weighting scheme. In other words, we convert each message in both datasets to a vector with N=2,400 dimensions. Next, we build the classification model using SVM as the classifier algorithm based on the generated vectors from the training dataset. Finally, using the built model we classify the vectors from the evaluation dataset and plot the results in Figure 3 by computing F1-score for classification in each month during the evaluation period of 2014. F1-score is a common measure of a test's accuracy, which considers both precision and recall of the test.

*Proposed approach (Word2Vec)*: As we describe in Section III, we apply the recently introduced Continuous Bag-of-Words and Skip-gram model [17]. Specifically, we take advantage of the pre-trained vectors published as part of its Word2Vec implementation[7]. For each word in a tweet we retrieve a corresponding vector from the Word2Vec dataset, where each vector has 300 dimensions. Since a tweet typically consists of multiple words, we compute a centroid vector based on all vectors retrieved for a given tweet [18]. Using this approach, we compute centroid vectors for the tweets in the training and evaluation datasets. Next, we build the classification model using SVM as the classifier algorithm based on the centroid vectors from the annotated training dataset. Finally, we use the built model to classify the centroid vectors in the evaluation dataset and show the F1-score results in the same Figure 3.

[7]https://code.google.com/p/word2vec/

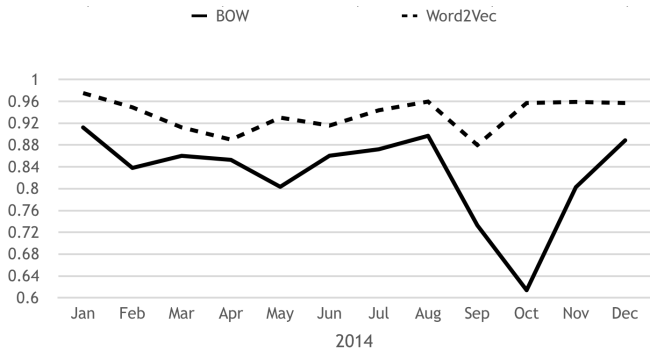


Fig. 3. Comparison of Word2Vec classification vs Bag-Of-Words

The proposed Word2Vec based classification approach consistently outperforms the baseline BOW approach in each month during the evaluation period. Note, that this result is achieved despite a significantly smaller number of features, which also makes the proposed method faster to execute than the baseline approach. The average F1-score achieved by the Word2Vec based classification approach during 2014 is 0.936 compared to 0.828 obtained by the BOW approach during the same period.

### C. Visualization of User Ranking

We draw a diagram of the relationships between users who post relevant tweets based on their retweet activity and show it in Figure 4. The diagram is implemented using D3.js, which is a JavaScript data visualization library [24]. Each node in the diagram represents a Twitter user and the links between users are based on their retweet relationships. The colors represent the number of relevant tweets a user has. The more relevant tweets the user has, the darker the corresponding node is. The size of a node is used to show the relevance ranking score computed by the system, such that the higher the score is, the bigger the node will be.

For visualization purposes, we add the labels of the users with the highest ranking scores next to their nodes. Most of higher relevance ranking users also have higher total rankings, such as *@YahooNews*. The center of each cloud is still the darker color node, which means the relevance ranking effectively represents the influence of the user. And since the relevance ranking is calculated using only relevant tweets, the relevance ranking also represents the relevance of the user.

As shown in Section IV, there are also users who post tweets that contain our search keywords, but whose meaning is irrelevant to our topic of interest, namely landslide as a disaster. The diagram of the relationships between users who

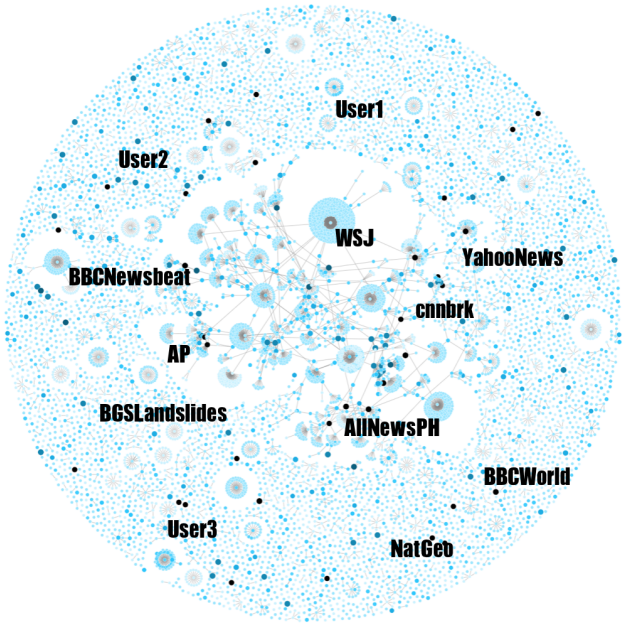


Fig. 4. Users posting relevant to landslide tweets

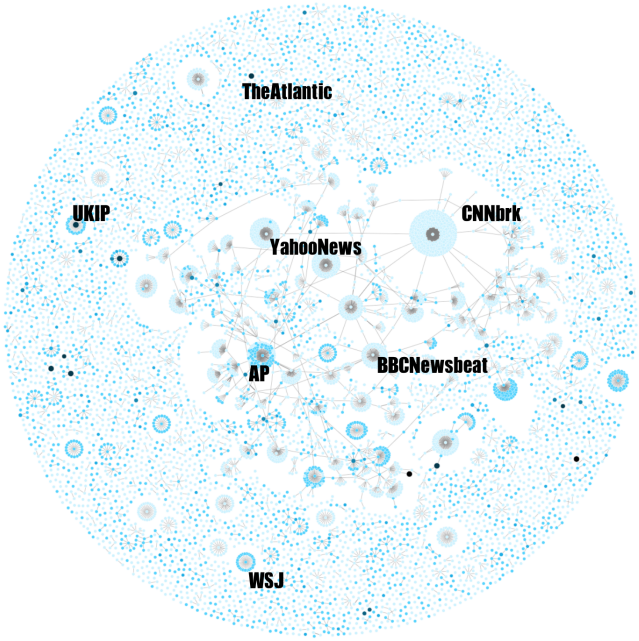


Fig. 5. Users posting irrelevant to landslide tweets

post irrelevant to disaster tweets is shown in Figure 5. The diagram is based on the same concepts, where each node is a Twitter user and the links between nodes are based on their retweet relationships. However, this time we use the tweets that discuss the topics, which are irrelevant to landslide as a disaster, as well as the users who posted or retweeted them. Again, we add the labels of the users with the highest ranking scores next to their nodes.

We can observe that multiple users are present in both diagrams. Most such users are authoritative news sources, including Wall Street Journal (*@WSJ*), BBC (*@BBCWorld* and *@BBCNewsbeat*) and Yahoo News (*@YahooNews*). Such news sources have Twitter accounts where they post messages containing links to the articles published in their online resources. Since these sources are authoritative, the messages they post attract attention in the form of retweets by other users. We want to score tweets posted by such users higher to reflect their influence among the corresponding community. Note, that a misclassified tweet by an influential user would negatively affect the detection results more than a misclassified tweet by a non-authoritative user based on this approach. However, authoritative users typically post messages about events that generate a considerable amount of public interest, so typically there is more than one tweet discussing such events. And since the proposed tweet classification system is highly accurate, then the majority of the tweets for each event will still be classified correctly, such that the events will still be correctly classified as demonstrated in the next experiment.

The proposed relevance ranking strategy can not only identify authoritative and relevant news sources but also successfully determine local users who actively post about landslides as natural disasters. There are local users with high relevance scores as described in Section IV. For example, with the proposed strategy, User1, a journalist, presenter, and local media partner, gets a relatively high relevance ranking score. Similarly, User2, a user posting information about the city of Timika and its surroundings, is identified as a relevant and influential local user, and User3, a lecturer by profession, coordinator, and chief editor, is labeled with high relevance ranking scores. Again, we replace the actual screen names of these users with anonymous labels to preserve their privacy.

### D. Evaluation of Event Detection

This experiment is designed to evaluate the performance of the proposed relevance ranking formula. Specifically, we compare the results of event detection based on the proposed formula shown in Equation (3) versus the existing approach based on Equation (2). We use the first three months of 2014 for evaluation purposes. For each cell we execute both formulas to detect the occurrence of a landslide based on the tweets mapped to that cell. The evaluation dataset contains a ground truth label for each cell that we use to evaluate the accuracy of the computed label. Figure 6 shows the F1-score of the accuracy of landslide detection based on the proposed relevance ranking formula versus the existing approach during each month in the evaluation period.

There is an increase in the F1-score based on the accuracy of the proposed strategy during January and February in 2014 compared to the F1-score of the existing method. Although the newly proposed strategy does not improve the F1-score compared to the existing approach in March 2014, but their values are relatively similar during that month. On average, the proposed relevance ranking strategy has an F1-score of 0.94 compared to 0.92 for the existing approach.

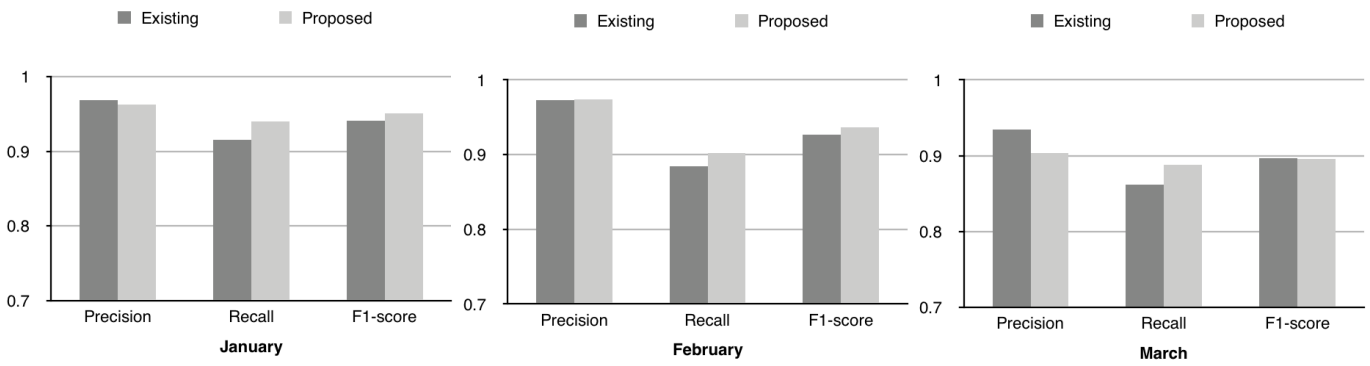


Fig. 6. Evaluation of the proposed relevance ranking strategy vs existing approach

## VII. Discussion

One of the important assumptions we make in this work is that all of the data items mapped to a given cell are related to the same event. For example, it is assumed that all tweets, which mention a particular place, discuss the same event. In reality, multiple events may be occurring in the same place, especially if the mentioned place is a large administrative unit, such as a district or a region. Such approach may lead to missed events, because an election held in a given place may attract more interest from the public than a local disaster, so that the system may incorrectly decide that there was no disaster in that place. Hence, the bigger the administrative unit of a given place the smaller the probability that all messages mentioning that place discuss the same event.

Hence, the proposed disaster information system may be improved further by taking into account the semantic meaning of the messages mapped to cells, such that similar messages are grouped into clusters within the cell and that the clusters are evaluated separately. Then a political event, even in a large administrative unit, would be correctly classified as an irrelevant event in that cell, while successfully detecting a disaster that occurred in the same place.

Another important assumption that we make is that only one disaster can occur within a 30-day window per cell. Although this is a very simplified approach, but it not only helps with the data analysis, but it also generally holds in most cases, such as landslide events.

Note, that the proposed system is fully automated and provides a comprehensive analysis of the data collected from social networking platforms. The system is comprehensive, because it analyzes every message collected from the social networking platforms. It also means that we can potentially compute the relevance ranking of each social user whose messages are available through the public data streams. Therefore, the system should be able to come up with a sound decision regarding detected events sooner as it will have more information about users beforehand.

Finally, we plan to improve the proposed relevance ranking formula further for several reasons. When we compute the relevance ranking score for a cell, the irrelevant and relevant user scores affect the result differently as we use the absolute values of those scores. We plan to normalize the values of the user relevance scores in the formula, e.g. by applying a common sigmoid normalization technique. This will also help to keep the cell scores within the [0, 1] range.

## VIII. Conclusion and Future Work

In this paper we provide an overview of a comprehensive disaster information system based on the data collected from social networking platforms. We use landslides as an illustrative disaster and briefly describe each step in the system's pipeline with the foci on the classification and ranking steps, which measure the relevance of individual messages as well as the groups of messages (or events) to disasters. We propose to build two virtual communities of users, namely "relevant" and "irrelevant", based on their relevance to landslide as a disaster and compute the users' influence in each of them. This allows us to improve the quality of the collected data on landslides by taking into account not only the messages posted by users, but also the users' influence and the amount of their activity in these communities. The proposed system achieves an average F1-score of 0.936 when classifying individual tweets and 0.941 when ranking the relevance of the events.

Our future work involves the evaluation of the proposed methods for a more comprehensive analysis of disasters based on multiple social networks, such as Twitter, Instagram, YouTube, and Facebook. Note, that disasters can be both natural and man-made, such as terrorist attacks. Early detection of emerging adverse events may help save lives and reduce the impact on society. Also, having comprehensive information related to the detected events may not only help during an actual crisis, but also afterwards for mitigation purposes.

## Acknowledgements

This research has been partially funded by National Science Foundation by CNS/SAVI (1250260, 1402266), IUCRC/FRP (1127904), CISE/CNS (1138666, 1421561) programs, and gifts, grants, or contracts from Fujitsu, HP, Intel, Singapore Government, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and

conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

## REFERENCES


[1] S.-H. Yang, A. Kolcz, A. Schlaikjer, and P. Gupta, "Large-scale high-precision topic modeling on Twitter," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1907–1916.

[2] B. L. Fraustino, Julia Daisy and Y. Jin, "Social media use during disasters: A review of the knowledge base and gaps," *Final Report to Human Factors/Behavioral Sciences Division, Science and Technology Directorate, U.S. Department of Homeland Security. College Park, MD: START*, 2012.

[3] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: What twitter may contribute to situational awareness," in *CHI '10*, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 1079–1088. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753486

[4] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles," in *CHI*, 2011.

[5] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *CIKM'10*, 2010.

[6] A. Musaev, D. Wang, and C. Pu, "LITMUS: A Multi-Service Composition System for Landslide Detection," *Services Computing, IEEE Transactions on*, vol. 8, no. 5, pp. 715–726, Sept 2015.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *WWW*, 2010.

[8] M. Guy, P. Earle, C. Ostrum, K. Gruchalla, and S. Horvath, "Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies," in *Proceedings of the 9th International Conference on Advances in Intelligent Data Analysis*, ser. IDA'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 42–53.

[9] A. Musaev, D. Wang, and C. Pu, "LITMUS: Landslide Detection by Integrating Multiple Sources," in *ISCRAM 2014 Conference Proceedings 11th International Conference on Information Systems for Crisis Response and Management*, 2014, pp. 677–686.

[10] A. Musaev, D. Wang, S. Shridhar, C.-A. Lai, and C. Pu, "Toward a real-time service for landslide detection: Augmented explicit semantic analysis and clustering composition approaches," in *Web Services (ICWS), 2015 IEEE International Conference on*. IEEE, 2015, pp. 511–518.

[11] A. Musaev, D. Wang, C. A. Cho, and C. Pu, "Landslide detection service based on composition of physical and social information services," in *Web Services (ICWS), 2014 IEEE International Conference on*, June 2014, pp. 97–104.

[12] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11. New York, NY, USA: ACM, 2011, pp. 45–54. [Online]. Available: http://doi.acm.org/10.1145/1935826.1935843

[13] P. Lawrence, B. Sergey, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Stanford University, Technical Report, 1998.

[14] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2, 1995, pp. 1137–1145.

[15] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[16] A. Musaev, D. Wang, S. Shridhar, and C. Pu, "Fast Text Classification Using Randomized Explicit Semantic Analysis," in *Information Reuse and Integration (IRI), 2015 IEEE International Conference on*. IEEE, 2015, pp. 364–371.

[17] T. Mikolov, W. tau Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," 2013.

[18] E.-H. S. Han and G. Karypis, *Principles of Data Mining and Knowledge Discovery: 4th European Conference, PKDD 2000 Lyon, France, September 13–16, 2000 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, ch. Centroid-Based Document Classification: Analysis and Experimental Results, pp. 424–431.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, 2009.

[20] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ser. WSDM '10. New York, NY, USA: ACM, 2010, pp. 261–270. [Online]. Available: http://doi.acm.org/10.1145/1718487.1718520

[21] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 363–370. [Online]. Available: http://dx.doi.org/10.3115/1219840.1219885

[22] Google Inc., "The Google Geocoding API," https://developers.google.com/maps/documentation/geocoding/, accessed on 4/1/2016.

[23] A. Musaev, D. Wang, and C. Pu, "Multi-hazard detection by integrating social media and physical sensors," in *Social Media for Government Services*. Springer, 2015, pp. 395–409.

[24] M. Bostock, V. Ogievetsky, and J. Heer, "$D^3$: Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.